THE T05 SYSTEM FOR THE VOICEMOS CHALLENGE 2024: TRANSFER LEARNING FROM DEEP IMAGE CLASSIFIER TO NATURALNESS MOS PREDICTION OF HIGH-QUALITY SYNTHETIC SPEECH

Kaito Baba, Wataru Nakata, Yuki Saito, Hiroshi Saruwatari

The University of Tokyo, Japan

ABSTRACT

We present our system (denoted as T05) for the VoiceMOS Challenge (VMC) 2024. Our system was designed for the VMC 2024 Track 1, which focused on the accurate prediction of naturalness mean opinion score (MOS) for high-quality synthetic speech. In addition to a pretrained self-supervised learning (SSL)-based speech feature extractor, our system incorporates a pretrained image feature extractor to capture the difference of synthetic speech observed in speech spectrograms. We first separately train two MOS predictors that use either of an SSL-based or spectrogram-based feature. Then, we fine-tune the two predictors for better MOS prediction using the fusion of two extracted features. In the VMC 2024 Track 1, our T05 system achieved first place in 7 out of 16 evaluation metrics and second place in the remaining 9 metrics, with a significant difference compared to those ranked third and below. We also report the results of our ablation study to investigate essential factors of our system.

***Index Terms*—** VMC 2024, MOS prediction, zoomed-in MOS test, SSL, feature fusion, deep image classifier

1. INTRODUCTION

Automatic quality assessment of synthetic speech is an emerging research topic in the text-to-speech (TTS) and voice conversion (VC) research fields [1, 2]. It is a promising technology for further development of TTS and VC because it can reduce the cost of human-based subjective evaluations on synthetic speech, such as a mean opinion score (MOS) test. In fact, UTMOS [3], an open-sourced MOS prediction system, was introduced as an alternative way to compare the performances of TTS systems submitted to the Interspeech 2024 Speech Processing Using Discrete Speech Unit Challenge [4]. Therefore, a MOS prediction system specialized for high-quality synthetic speech is valuable for a unified comparison of state-of-the-art deep neural network (DNN)-based TTS/VC systems [5].

The range-equalizing bias in MOS tests [6] is one challenge to be addressed for achieving this goal. That is, listeners in a MOS test tend to use the entire range of choices on the rating scale (e.g., from one to five), regardless of the absolute quality of the samples used in the MOS test. For example, a medium-quality TTS/VC system in one MOS test may achieve relatively low MOS in another test excluding worse-performing systems from the comparison (i.e., zoomed-in MOS test). Therefore, MOS prediction systems built without considering the range-equalizing bias may underestimate high-quality synthetic speech or overestimate low-quality synthetic speech.

In this paper, we present our MOS prediction system specialized for high-quality synthetic speech, which is designed for the VoiceMOS Challenge (VMC) 2024 [7] Track 1, the task of predicting zoomed-in MOS test results. Our system adopts some techniques that can improve the MOS prediction performance in the VMC 2022 and 2023 [8, 9]: using self-supervised learning (SSL)-based speech features [3] and fusing multiple speech features [10]. We also investigate the effectiveness of using EfficientNetV2 [11], i.e., DNN-based *image* feature extractor, for capturing the difference of synthetic speech observed in speech spectrograms accurately. In our two-stage fine-tuning strategy, we first separately train two MOS predictors that use either of an SSL-based or spectrogram-based feature. Then, we fine-tune the two predictors for better MOS prediction using the fusion of two extracted features. In the VMC 2024 Track 1, our T05 system achieved first place in 7 out of 16 evaluation metrics and second place in the remaining 9 metrics, with a significant difference compared to those ranked third and below. We also report the results of our ablation study to investigate essential factors of our systems. The result demonstrates that fusing the two features improves the correlation-based evaluation metrics. It also indicates that using a large-scale MOS dataset consisting of solely neural TTS samples or an actual zoomed-in MOS dataset for the training enhances the MOS prediction performance. The code and the demo for our system are available online[1].

2. THE VMC 2024 TRACK 1

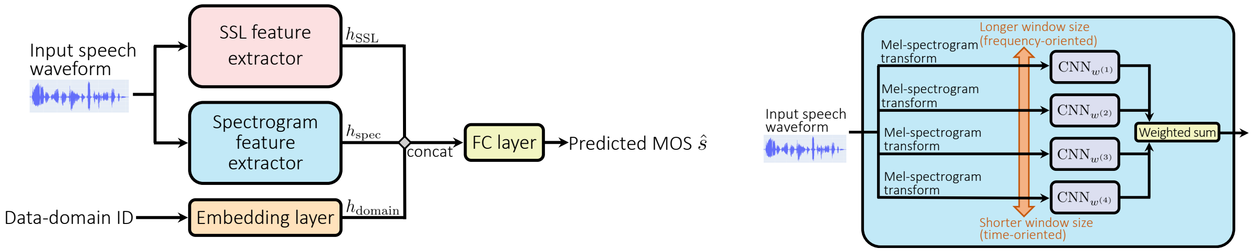
The VMC 2024 [7] consists of three tracks, where our T05 system is designed for the Track 1. In this track, the organizers collected the results of zoomed-in MOS tests, where they compared speech synthesis systems that achieved high MOS from the BVCC dataset [12]. The organizer conducted three MOS tests with the zoom-in rates of 50%, 25%, and 12%, representing the number of systems covered in the test compared to the original BVCC dataset. No official training data considering these "zoomed-in" situations were provided by the organizers, and thus participants were required to build their MOS prediction systems with publicly available MOS datasets. After the track finished, the organizers disclosed that the validation set consisted of the results of 50% zoomed-in MOS test, while the evaluation set consisted of both 25% and 12% zoomed-in MOS tests. The evaluation metrics included mean squared error (MSE), linear correlation coefficient (LCC), Spearman's rank correlation coefficient (SRCC), and Kendall's rank correlation coefficient (KTAU) at both the utterance and system levels.

3. OUR SUBMITTED SYSTEM (UTMOSV2)

3.1. Basic Architecture

Our T05 system (UTMOSv2) leverages the combination of spectrogram features extracted by a pretrained image feature extractor and speech features obtained from pretrained speech SSL models (i.e., SSL feature). Figure 1a illustrates the basic model architecture.

[1] Code: https://github.com/sarulab-speech/UTMOSv2
Demo: https://huggingface.co/spaces/sarulab-speech/UTMOSv2



(a) Basic architecture.

(b) Spectrogram feature extractor.

Fig. 1: The basic model architectures in the proposed system (UTMOSv2). Our system leverages SSL features $\boldsymbol{h}_{\mathrm{SSL}}$ and spectrogram features $\boldsymbol{h}_{\mathrm{spec}}$. Additionally, the data domain embedding $\boldsymbol{h}_{\mathrm{domain}}$ is obtained from the data-domain ID which is unique to each dataset used in the training. Finally, these three features are concatenated to predict the MOS of the input speech.

3.1.1. Spectrogram Feature Extractor

The field of computer vision using deep learning has significantly advanced in recent years, and applying DNN-based models (i.e., deep image classifiers) to spectrograms has demonstrated promising results in some audio/speech processing tasks [13, 14, 15]. Our system thus leverages the features extracted from spectrograms using a convolutional neural network (CNN) pretrained on a large image dataset. The architecture of our spectrogram feature extractor is shown in Figure 1b.

In spectrogram feature extraction, the input speech waveform is first transformed into multiple mel-spectrograms. Each mel-spectrogram is extracted with different short-term Fourier transform (STFT) settings, which aims to mitigate the problem of the trade-off between frequency resolution and time resolution determined by the window size [13]. Let $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_K^\top]^\top$ be K speech frames, where $\boldsymbol{x}_k$ denotes the kth frame consisting of L samples. These frames are randomly extracted from the input speech waveform. Multiple mel-spectrogram transformations with N different window sizes $(w^{(1)}, \cdots, w^{(N)})$, $w^{(n)} \in \mathbb{N}$, $\mathrm{MelSpec}_{w^{(n)}}(\cdot)$, are applied to each extracted audio frame:

$$\boldsymbol{y}_k^{(n)} = \mathrm{MelSpec}_{w^{(n)}}(\boldsymbol{x}_k),$$

where $\boldsymbol{y}_k^{(n)}$ denotes the nth mel-spectrogram extracted from the kth speech frame using the window size $w^{(n)}$.

These mel-spectrograms are then regarded as images rather than speech parameter sequences and fed into CNNs pretrained on ImageNet [16], following previous work [15]. The shape of mel-spectrogram image is fixed as (F, F), where F represents the number of mel-bands, regardless of the window size setting. Multiple CNNs are prepared, where each network receives a spectrogram with a different window size $w^{(n)}$ as input, extracting an image feature as follows:

$$\boldsymbol{h}_k^{(n)} = \mathrm{CNN}_{w^{(n)}}(\boldsymbol{y}_k^{(n)}).$$

The features obtained from $\boldsymbol{y}_k^{(n)}$ through the CNN for each window width setting, i.e., $(\boldsymbol{h}_k^{(1)}, \ldots, \boldsymbol{h}_k^{(N)})$, are aggregated using a weighted sum $\tilde{\boldsymbol{h}}_k = \sum_{n=1}^{N} w_{\mathrm{spec},n} \boldsymbol{h}_k^{(n)}$. The trainable weight parameter vector $\boldsymbol{w}_{\mathrm{spec}} \in \mathbb{R}^N$ is initialized such that $\sum_{n=1}^{N} w_{\mathrm{spec},n} = 1$. As a result, the aggregated feature $\tilde{\boldsymbol{h}}_k$ has the dimension $\mathbb{R}^{c \times f \times t}$, where c, f, and t denote the number of features, the height of feature maps and the width of the feature maps obtained through CNNs, respectively. These aggregated features with different k are then concatenated across several frames in the t dimension and subsequently pooled in both t and f dimensions. A combination of average and max pooling is used in the time direction; a combination of attention [17] and max pooling was employed in the frequency direction. The final output of our spectrogram feature extractor is hereinafter denoted as $\boldsymbol{h}_{\mathrm{spec}}$.

3.1.2. SSL Feature Extractor

Following previous studies on automatic MOS prediction [2, 3], we utilize a pretrained SSL model to extract speech features from an input waveform. The raw waveform is first fed into the SSL model to extract hidden states from the each layer of the Transformer encoder $(\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M)$. Then, the hidden states are aggregated using a weighted sum $\tilde{\boldsymbol{e}} = \sum_{m=1}^{M} w_{\mathrm{SSL},m} \boldsymbol{e}_m$, where M denotes the number of Transformer encoder layers. The trainable weight parameter vector $\boldsymbol{w}_{\mathrm{SSL}} \in \mathbb{R}^M$ is initialized such that $\sum_{m=1}^{M} w_{\mathrm{SSL},m} = 1$. Finally, unlike in previous studies [2, 3], combination of attention [17] and max pooling along the sequence dimension are applied to the aggregated hidden state vectors for each time step. The final output of our SSL feature extractor is hereinafter denoted as $\boldsymbol{h}_{\mathrm{SSL}}$.

3.1.3. Data-domain Encoding

Following the UTMOS system [3], we build our MOS prediction system using multiple MOS datasets for the model training with the data-domain encoding (i.e., conditioning the system on the dataset ID). This aims to address the biases in different MOS tests, possibly including the range-equalizing bias [6]. For the data-domain encoding, simple look-up embedding table is used for converting discrete dataset ID to continuous data-domain embedding $\boldsymbol{h}_{\mathrm{domain}}$.

Note that this data-domain encoding cannot define IDs for unseen MOS datasets and thus does not necessarily work properly for the out-of-domain prediction. One can deal with this issue by, for example, predicting MOS for some seen data-domains and taking the average of multiple predicted scores [3]. Because the primal focus of this paper is the range-equalizing bias, we thoroughly investigate the domain gap between the training and test datasets in our ablation study (Section 4.6).

3.1.4. Fusion of Spectrogram Features and SSL Features

A simple fully connected layer is prepared and trained for predicting the MOS of input speech using the fusion of extracted spectrogram and SSL features denoted as $\boldsymbol{h}_{\mathrm{spec}}$ and $\boldsymbol{h}_{\mathrm{SSL}}$, respectively. The input is the concatenation of these two features and the data-domain embedding along the feature dimension:

$$\hat{s} = \mathrm{FC}\left(\mathrm{Concat}(\boldsymbol{h}_{\mathrm{spec}}, \boldsymbol{h}_{\mathrm{SSL}}, \boldsymbol{h}_{\mathrm{domain}})\right), \tag{1}$$

where $\mathrm{FC}(\cdot)$ and $\mathrm{Concat}(\cdot)$ denote a fully connected layer and feature concatenation, respectively.

Table 1: Specifications of dataset used in the training. "BC" and # mean "Blizzard Challenge" and "number of," respectively.

Dataset name	# listeners	# systems	# sentences	# ratings
BC2008	229	7	80	16,987
BC2009	129	19	141	21,332
BC2010 EH1	177	18	36	5,863
BC2010 EH2	179	18	36	6,070
BC2010 ES1	73	8	16	1,152
BC2010 ES3	84	8	16	1,250
BC2011	236	13	39	9,328
BVCC	304	187	7106	56,848
SOMOS	987	201	2000	359,100
sarulab-data	304	95	3610	28,880

3.2. Additional Data Collection

As there are no official training sets provided by the organizers, we collected training data from the publicly available MOS test results. The collected data consisted of BVCC [12], Blizzard Challenge (BC) 2008 [18], 2009 [19], 2010 [20], 2011 [21] SOMOS [22], and zoomed-in BVCC dataset that is publicly available (sarulab-data)[2]. The specification of datasets are shown in Table 1.

For the dataset derived from BC, we only used subjective evaluation results for the english utterances. For BC2008, We excluded listeners which are marked with EUS as their scores were not in 5-point scale. For BC2010, We used results for task EH1, EH2, ES1 and ES3. ES2 was excluded as naturalness of synthetic speech was not considered in this task.

3.3. Loss Function

For the loss function used in the training, we adopt the combination of a contrastive loss [3] and mean squared error (MSE) loss. Specifically, the contrastive loss is formulated as

$$\mathcal{L}_{\text{con}}(s, \hat{s}) = \sum_{i \neq j} \max(0, |(s_i - s_j) - (\hat{s}_i - \hat{s}_j)| - \alpha), \quad (2)$$

where s and $\hat{s}$ denote the target MOS and predicted MOS, respectively. The margin hyperparameter $\alpha > 0$ makes the trained model ignore small errors lower than this margin. The final loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L}(s, \hat{s}) = \lambda_{\text{con}} \mathcal{L}_{\text{con}}(s, \hat{s}) + \lambda_{\text{mse}} \mathcal{L}_{\text{mse}}(s, \hat{s}), \quad (3)$$

where λ_{con} and λ_{mse} are hyperparameters that control the weights of the contrastive and MSE loss functions, respectively.

3.4. Multi-Stage Learning

When fine-tuning a pretrained model, catastrophic forgetting can significantly worsen the performance of the model on learned domains [23]. To mitigate this, we introduce multi-stage learning.

Since our proposed system is large and difficult to train the parameters of two feature extractors from scratch, we first train the two extractors separately. Then, we fine-tune the pretrained weights from these individual models and train the parameters of the FC layer (Eq. 1) for the feature fusion. In summary, the training performs the following stages:

Stage 1: The spectrogram and SSL feature extractors are trained separately. Specifically, an FC layer, which takes the concatenated features of data-domain embedding $\boldsymbol{h}_{\text{domain}}$ and either of $\boldsymbol{h}_{\text{spec}}$ or $\boldsymbol{h}_{\text{SSL}}$ and predicts MOS, is trained jointly with the extractor.

Stage 2: The weights of the two extractors are frozen and only the feature fusion layer (Eq. 1) along with a new data-domain embedding layer is trained.

Stage 3: All parameters of the models in our system are fine-tuned with a small learning rate.

Our SSL feature extractor is also pretrained with two-stage training following similar stages described above. That is, the model parameters of a backbone SSL model is first frozen and only the FC layer for the MOS prediction is trained. Then, all parameters of this extractor including the SSL model are fine-tuned. In contrast, our preliminary experiment showed that this two-stage pretraining for the spectrogram feature extractor did not bring significant improvement. Therefore, we decided to train the entire model of this extractor, i.e., the pretrained CNNs and the FC layer for the MOS prediction.

Technically, in the comparative experiments in Section 4, the spectrogram feature extractor was trained using data-domain embeddings on all datasets. Meanwhile, the system submitted for the VMC2024 Track 1 excluded the data-domain encoding and performed fine-tuning on sarulab-data after the training on BVCC. Apart from this aspect, the DNN architecture used in the comparative experiments in Section 4 and the submitted system is exactly the same.

4. EXPERIMENTS

We conducted several experiments to validate the effectiveness of our T05 system. Specifically, we performed ablation studies on the fusion of spectrogram and SSL features, multi-stage learning, and datasets.

4.1. Common Experimental Conditions

We used EfficientNetV2 [24] as the CNN for our spectrogram feature extractor. For the backbone SSL model, we used wav2vec2.0 [25] base[3] pretrained on LibriSpeech [26]. For the data-domain encoding, we used embedding with hidden size of 1.

For the loss function, we set the margin hyperparameter $\alpha = 0.2$ (Eq. (2)) for all experiments. The weight coefficients for the contrastive and MSE loss, λ_{con} and λ_{mse}, were set to of 0.2 and 0.7, respectively. These hyperparameters were decided based on our preliminary experiments. For the optimizer, we used AdamW [27] with the weight decay coefficient of 1×10^{-4}. For learning rate scheduler, we decayed the learning rate with a cosine annealing [28]. The initial learning rate varied depending on the training stage. During training, we incorporated mixup [29] for all training, which was shown to be effective in MOS prediction [30].

A five-fold cross-validation was performed, and the best model checkpoint was selected based on the average system-level SRCC calculated for each validation fold. The final prediction was obtained by averaging the predictions from each of the five folds. Additionally, during inference, we generated predictions five times by randomly selecting different frames of the input speech waveform and then averaged these predictions (i.e., test-time augmentation [31]).

[2] https://github.com/sarulab-speech/VMC2024-sarulab-data

[3] https://huggingface.co/facebook/wav2vec2-base

Table 2: Comparison of performance between our systems and the "B01" baseline. **Bold** and underlined scores are the best and worst among our three systems, respectively. We also compare our systems and human-annotated MOS ("BVCC MOS").

	Zoom-in rate: 25%								Zoom-in rate: 12%							
	Utterance-level				System-level				Utterance-level				System-level			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
Ours	0.690	**0.618**	**0.613**	**0.442**	0.465	**0.922**	**0.919**	**0.752**	0.459	**0.578**	**0.579**	**0.404**	0.288	**0.840**	**0.854**	**0.650**
w/o SSL	**0.566**	0.576	0.565	0.403	**0.353**	0.889	0.909	0.740	**0.357**	0.518	0.516	0.355	**0.188**	0.762	0.770	0.570
w/o spec.	0.937	0.602	0.603	0.432	0.700	0.910	0.909	0.731	0.673	0.530	0.529	0.364	0.497	0.793	0.793	0.570
B01	1.154	0.508	0.509	0.358	0.998	0.750	0.745	0.539	0.741	0.422	0.417	0.285	0.589	0.608	0.609	0.444
UTMOS [3]	0.872	0.407	0.411	0.286	0.690	0.649	0.615	0.433	0.541	0.297	0.300	0.206	0.378	0.440	0.367	0.230
BVCC MOS	0.717	0.377	0.358	0.256	0.413	0.728	0.679	0.495	0.481	0.322	0.316	0.225	0.223	0.691	0.702	0.467

4.2. Evaluation Metrics

The evaluation was performed on the test set with the zoom-in rate of 25% and 12%. In both test sets, we used system level and utterance-level MSE, LCC, SRCC and KTAU as metrics, referring to the VMC2024 evaluation protocol.

4.3. VMC2024 Results of Our T05 System [7]

In the Track1, both utterance-level and system-level metrics are calculated for 25% and 12% highest-rated systems, respectively. The official evaluation results show that our T05 system achieved the first place in 7 out of 16 metrics and ranked the second in the remaining 9 metrics, thereby securing either the first or second place in all metrics. Additionally, it is notable that there is a large margin in the performance to those ranked the third and below.

4.4. Ablation Study on Fusing Spectrogram/SSL Features

To evaluate the effectiveness of fusing spectrogram features and SSL features, we compared the prediction scores of the fused model with those obtained using only spectrogram or SSL features.

4.4.1. Experimental Conditions

In this ablation study, we compared the following systems:

- **Ours**: The proposed system using the feature fusion.
- **Ours w/o SSL**: The proposed system using only the spectrogram feature extractor.
- **Ours w/o spec.**: The proposed system using only the SSL feature extractor.
- **B01:** SSL-MOS [2] trained on the original BVCC [12] samples and labels. This system was considered as baseline system in the VMC 2024 track 1.
- **UTMOS [3]:** The opensourced MOS prediction system.

"Ours w/o SSL" was trained with a learning rate ranging from 1×10^{-3} to 1×10^{-7}, a batch size of 10, and for 20 epochs. As explained in Section 3.4, "Ours w/o spec." was built with the two-stage training. We first trained the FC layer and data-domain embedding for 20 epochs using the learning rate ranging from 1×10^{-3} to 1×10^{-7} and batch size of 32. Then, we fine-tuned all model parameters for 5 epochs using the learning rate ranging from 3×10^{-5} to 1×10^{-9} and batch size of 32. The system using the feature fusion, "Ours," was built upon these two systems. Specifically, we utilized these two feature extractors trained through "Ours w/o spec." and "Ours w/o SSL." The following FC layer, and data-domain embedding were randomly initialized.

The stage 2 training was performed for 8 epochs using a learning rate ranging from 1×10^{-3} to 1×10^{-5} and a batch size of 16. The stage 3 training was iterated with 2 epochs using a learning rate ranging from 5×10^{-5} to 1×10^{-8} and a batch size of 8.

In this comparison, we used all datasets listed in Table 1 for the training and set the data-domain ID for the MOS prediction to "BVCC" in three our systems.

4.4.2. Results and Discussion

The results are shown in Table 2. For correlation-based metrics, we can see that all of our three systems consistently outperforms both two baseline models in all metrics. Furthermore, in correlation-based metrics, while there are little difference in scores between "Ours w/o SSL" and "Ours w/o spec.," the fusion system, "Ours," demonstrates a significant improvement in scores compared to these two systems. These results indicate that our systems are more effective for zoomed-in MOS prediction compared to the existing baseline systems, particularly in correlation-based metrics. It also suggests the effectiveness of fusing spectrogram and SSL features in these metrics.

One noteworthy observation is that "Ours w/o SSL" achieves the best MSE in all cases, but the worst in many cases in the correlation-based metrics. On the other hand, "Ours w/o spec." scored the highest MSE, but outperforms "Ours w/o SSL" in many cases in the correlation-based metrics. From this perspective, we can infer that the spectrogram features derived from our image feature extractor are better at capturing fine differences in synthetic speech and predicting absolute MOS values, while SSL features are better at predicting rankings among multiple speech synthesis systems. In summary, these results suggest that the fusion of these features improves the prediction of absolute speech quality while further improving the correlation-based measures.

We also computed the evaluation metrics between the ground-truth MOS and human-annotated MOS ("BVCC MOS"), which was collected without considering the range-equalizing bias. The results from Table 2 demonstrate that the bias actually exists and "BVCC MOS" is not well correlated with the ground-truth MOS. In contrast, our fusion system shows better scores than "BVCC MOS" in all metrics except for system-level MSE. Considering that the prediction is made with the data-domain embedding of BVCC, these results suggest that our system has demonstrated robust prediction of MOS for unseen listening test settings.

4.5. Comparison of Multi-Stage Learning

To evaluate the effectiveness of the multi-stage learning described in Section 3.4, we conducted a comparative experiment.

Table 3: Comparison of performance between our systems that did not employ the multi-stage learning process. **Bold values** are the best scores and underlined values are the worst scores among each column.

	Zoom-in rate: 25%								Zoom-in rate: 12%							
	Utterance-level				System-level				Utterance-level				System-level			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
Ours	0.690	**0.618**	**0.613**	**0.442**	0.465	**0.922**	**0.919**	**0.752**	0.459	**0.578**	**0.579**	**0.404**	0.288	**0.840**	**0.854**	**0.650**
w/o Stage 2	0.469	0.531	0.555	0.394	0.209	0.900	0.911	0.744	0.342	0.436	0.505	0.350	0.108	0.787	0.816	0.602
w/o Stage 1&2	**0.355**	0.480	0.482	0.336	**0.125**	0.738	0.710	0.499	**0.293**	0.421	0.423	0.289	**0.097**	0.675	0.672	0.531

4.5.1. Experimental Conditions

In this experiment, we compared "Ours" in Section 4.4 with the following systems:

- **Ours w/o Stage 2**: The proposed system without performing the stage 2 training.
- **Ours w/o Stage 1&2**: The proposed system with performing only the stage 3 training.

The fine-tuning for "Ours w/o Stage 2" was performed for 20 epochs using a learning rate ranging from 1×10^{-4} to 1×10^{-7} and batch size of 8. The training for "Ours w/o Stage 1&2" ran 20 epochs using a learning rate ranging from 1×10^{-3} to 1×10^{-7} and batch size of 8. The training dataset and target domain-ID setting was the same as those used in Section 4.4.

4.5.2. Results and Discussion

The results are shown in Table 3. As the number of multi-stage learning stages is reduced and the two feature extractors are no longer pre-trained for the MOS prediction task, the behavior of the learned models can be seen to approach "Ours w/o SSL" (i.e., lower MSE and lower correlation-based metrics). This may be desirable in situations where we want to accurately predict the absolute MOS, but not when we want to compare different speech synthesis systems. In summary, these results suggest that the proposed multi-stage learning is essential for boosting the ability of the SSL features to capture differences between multiple synthetic speech samples.

This might be because the SSL and spectrogram were combined and trained before being optimized individually. Due to the different learning speeds of SSL feature extractor and the spectrogram feature extractor, the fully connected layer might have resulted in a model that emphasizes one over the other. Specifically, in this case, the spectrogram features might have been given more importance, leading the system to resemble "Ours w/o SSL."

4.6. Investigation on Dataset

To investigate which datasets described in Section 3.2 were effective for predicting the MOS for the zoomed-in target, i.e., newly obtained through listening tests of BVCC's top-performing systems, we conducted ablation studies on these datasets.

4.6.1. Experimental Conditions

For predicting MOS, we used "Ours" built with the almost same experimental setting as described in Section 4.4. Here, we changed the datasets for the training and the data-domain ID for the inference. Specifically, we trained "Ours" with "All datasets" and that without {BVCC, BC, SOMOS, sarulab-data}. This experiment enabled us to examine which dataset was essential for improving the MOS prediction performance in the zoomed-in test situation. In addition, by examining the prediction results when changing the data-domain ID, we can verify which domain (i.e. dataset) was closer to the zoomed-in dataset used in the VMC2024 Track 1. Note that only the mean values are presented in the results for the BC datasets, even though the data-domain ID was prepared for each BC dataset.

4.6.2. Results and Discussion

The results are shown on Table 4. In terms of the training datasets, "All datasets" achieves the best scores. However, in some cases the scores improve by excluding BVCC or BC from the training data. In addition, excluding SOMOS or sarulab-data from the training data tends to degrade the MOS prediction performance significantly. These results suggest that when building MOS prediction systems to compare the performance of high-quality speech synthesis, it is crucial to exclude datasets that are likely to contain low-quality speech systems when training. They also indicate that using MOS datasets containing as many results as possible from evaluation of synthetic speech produced by state-of-the-art DNN-based speech synthesis.

Focusing on the difference among data-domain for the MOS prediction, the MSE is the lowest for sarulab-data (i.e., the 50% zoomed-in BVCC) and the highest for BVCC, which clearly shows the effect of range-equalizing bias [6]. However, this tendency is not observed when comparing the correlation-based metrics. These results suggest that the negative effects caused by the range-equalizing bias are dominant in the prediction of the absolute MOS.

Additionally, when comparing the scores of correlation-based metrics between datasets with a 25% zoomed-in rate and those with a 12% zoomed-in rate, it can be observed that the scores are better for the 25% zoomed-in rate datasets in almost all cases. This suggests that the quality of the speech data used for training was closer to that of the 25% zoomed-in rate datasets.

5. CONCLUSION

In this paper, we presented our automatic MOS prediction system (UTMOSv2) submitted to the VMC 2024. Our system achieved first place in 7 out of 16 metrics in the VMC 2024 Track 1. The submitted T05 system leverages the fusion of spectrogram features from a pretrained image feature extractor and speech features from pretrained speech SSL models. Additionally, multi-stage learning and the use of multiple datasets were introduced. In the ablation study, we demonstrated that combining spectrogram features and SSL features improves the correlation-based metrics, while the MSE was best when only the spectrogram feature was used. Furthermore, the use of a wider range of datasets and multi-stage learning enhanced the performance of the MOS prediction. Future work includes constructing a MOS prediction system not only for the naturalness of synthetic speech but also for other aspects of speech, such as prosody.

6. ACKNOWLEDGEMENTS

This work was supported by JST Moonshot JPMJMS2011. We are also grateful to Kazuki Yamauchi, Osamu Take, and Takaki Hamada of the University of Tokyo for their helpful discussions.

Table 4: Results for ablation study regarding the training datasets and dataset domains. For example, the second-through-fifth columns list the MOS prediction performance for the VMC2024 Track 1 evaluation set using "BVCC" as the data-domain. Values in **bold face** shows the best result in each column and the underlined values show the worst result in each columns. Only the mean values are presented for the BC datasets.

(a) Utterance-level results at 25% zoomed-in rate.

	BVCC				BC				SOMOS				sarulab-data			
Training datasets	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
All datasets	0.690	**0.618**	**0.613**	**0.442**	**0.398**	0.620	0.616	0.444	**0.326**	0.617	0.612	0.441	**0.279**	0.620	0.615	0.444
w/o BVCC	–	–	–	–	0.741	**0.668**	**0.656**	**0.480**	0.444	**0.668**	**0.656**	**0.480**	0.386	**0.667**	**0.655**	**0.479**
w/o BC	**0.569**	0.531	0.533	0.378	–	–	–	–	0.414	0.528	0.527	0.374	0.329	0.543	0.530	0.377
w/o SOMOS	0.683	0.417	0.411	0.286	0.677	0.417	0.410	0.286	–	–	–	–	0.678	0.416	0.408	0.285
w/o sarulab-data	0.733	0.473	0.470	0.329	0.438	0.475	0.473	0.332	0.592	0.474	0.470	0.330	–	–	–	–

(b) System-level results at 25% zoomed-in rate.

	BVCC				BC				SOMOS				sarulab-data			
Training datasets	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
All datasets	0.465	**0.922**	**0.919**	**0.752**	**0.167**	**0.924**	**0.919**	**0.754**	**0.092**	**0.924**	0.916	0.744	**0.044**	**0.923**	**0.921**	**0.756**
w/o BVCC	–	–	–	–	0.487	0.910	0.918	0.748	0.194	0.911	**0.918**	**0.748**	0.138	0.910	0.919	0.752
w/o BC	**0.284**	0.912	0.916	0.746	–	–	–	–	0.133	0.885	0.891	0.725	0.051	0.899	0.911	0.744
w/o SOMOS	0.423	0.714	0.669	0.474	0.415	0.714	0.664	0.472	–	–	–	–	0.416	0.708	0.655	0.464
w/o sarulab-data	0.484	0.750	0.718	0.516	0.179	0.753	0.717	0.518	0.338	0.755	0.720	0.520	–	–	–	–

(c) Utterance-level results at 12% zoomed-in rate.

	BVCC				BC				SOMOS				sarulab-data			
Training datasets	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
All datasets	0.459	**0.578**	**0.579**	**0.404**	**0.262**	0.584	0.584	0.408	**0.234**	0.579	0.579	0.403	**0.238**	0.581	0.582	0.406
w/o BVCC	–	–	–	–	0.541	**0.633**	**0.626**	**0.452**	0.324	**0.636**	**0.629**	**0.454**	0.297	**0.636**	**0.629**	**0.454**
w/o BC	**0.393**	0.471	0.473	0.330	–	–	–	–	0.299	0.491	0.493	0.343	0.360	0.442	0.450	0.313
w/o SOMOS	0.447	0.369	0.376	0.257	0.443	0.370	0.375	0.256	–	–	–	–	0.443	0.370	0.378	0.258
w/o sarulab-data	0.484	0.429	0.430	0.293	0.312	0.427	0.431	0.293	0.392	0.427	0.428	0.292	–	–	–	–

(d) System-level results at 12% zoomed-in rate.

	BVCC				BC				SOMOS				sarulab-data			
Training datasets	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
All datasets	0.288	**0.840**	**0.854**	**0.650**	**0.088**	**0.844**	**0.851**	0.650	**0.056**	**0.840**	0.844	0.642	**0.058**	**0.842**	**0.838**	0.634
w/o BVCC	–	–	–	–	0.343	0.823	0.832	**0.665**	0.128	0.824	**0.846**	**0.681**	0.101	0.825	0.836	**0.673**
w/o BC	**0.145**	0.826	0.819	0.610	–	–	–	–	0.069	0.804	0.823	0.642	0.122	0.756	0.805	0.602
w/o SOMOS	0.224	0.667	0.696	0.467	0.221	0.665	0.682	0.459	–	–	–	–	0.221	0.665	0.700	0.483
w/o sarulab-data	0.282	0.671	0.647	0.448	0.102	0.674	0.661	0.459	0.186	0.675	0.690	0.483	–	–	–	–

7. REFERENCES

[1] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[2] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of mos prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[3] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.

[4] Xuankai Chang, Jiatong Shi, Jinchuan Tian, Yuning Wu, Yuxun Tang, Yihan Wu, Shinji Watanabe, Yossi Adi, Xie Chen, and Qin Jin, "The Interspeech 2024 Challenge on Speech Processing Using Discrete Units," in *Proc. Interspeech*, 2024, pp. 2559–2563.

[5] Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao, "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. ICML*, 2024.

[6] Erica Cooper and Junichi Yamagishi, "Investigating range-equalizing bias in mean opinion score ratings of synthesized speech," in *Proc. Interspeech*, 2023, pp. 1104–1108.

[7] Wen-Chin Huang, Szu-Wei Fu, Erica Cooper, Ryandhimas E. Zezario, Tomoki Toda, Hsin-Min Wang, Junichi Yamagishi, and Yu Tsao, "The VoiceMOS Challenge 2024: Beyond speech quality prediction," in *Proc. SLT*, 2024.

[8] Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech*, 2022, pp. 4536–4540.

[9] Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, "The VoiceMOS Chal-

lenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *Proc. ASRU*, 2023.

[10] Zili Qi, Xinhui Hu, Wangjin Zhou, Sheng Li, Hao Wu, Jian Lu, and Xinkang Xu, "LE-SSL-MOS: Self-supervised learning MOS prediction with listener enhancement," in *Proc. ASRU*, 2023.

[11] Mingxing Tan and Quoc V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. ICML*, 2021.

[12] Erica Cooper and Junichi Yamagishi, "How do voices from past speech synthesis challenges compare today?," in *Proc. SSW*, 2021, pp. 183–188.

[13] Jeno Szep and Salim Hariri, "Paralinguistic classification of mask wearing by image classifiers and fusion," in *Proc. Interspeech*, 2020, pp. 2087–2091.

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[15] Shahin Amiriparian, Nicholas Cummins, Sandra Ottl, Maurice Gerczuk, and Björn Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proc. ACIIW*, 2017, pp. 26–29.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009, pp. 248–255.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[18] Vasilis Karaiskos, Simon King, Robert AJ Clark, and Catherine Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, 2008.

[19] Alan W Black, Simon King, and Keiichi Tokuda, "The Blizzard Challenge 2009," in *Proc. Blizzard Challenge Workshop*, 2009, pp. 1–24.

[20] Alan W Black, Simon King, and Keiichi Tokuda, "The Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010.

[21] Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2011," in *Proc. The Blizzard Challenge Workshop*, 2011, pp. 1–10.

[22] Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Proc. Interspeech*, 2022, pp. 2388–2392.

[23] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, Jun. 1999.

[24] Mingxing Tan and Quoc V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. ICML*, 2021.

[25] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, vol. 33, pp. 12449–12460.

[26] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

[27] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[28] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. ICLR*, 2017.

[29] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[30] Kexin Wang, Yunlong Zhao, Qianqian Dong, Tom Ko, and Mingxuan Wang, "MOSPC: MOS prediction based on pairwise comparison," in *Proc. ACL*, 2023, pp. 1547–1556.

[31] Divya Shanmugam, Davis W. Blalock, Guha Balakrishnan, and John V. Guttag, "Better aggregation in test-time augmentation," in *Proc. ICCV*, 2021, pp. 1194–1203.